
A Principled Infotheoretic ϕ -like Measure

Virgil Griffith

Computation and Neural Systems, Caltech, Pasadena, CA 91125
i@virgil.gr

Abstract

Integrated information theory [1–3] is a mathematical, quantifiable theory of conscious experience. The linchpin of this theory, the ϕ measure, quantifies a system’s irreducibility to disjoint parts. Purely as a measure of irreducibility, we pinpoint three concerns about ϕ and propose a revised measure, ψ , which addresses them. Our measure ψ is rigorously grounded in Partial Information Decomposition and is faster to compute than ϕ .

1 Introduction

The measure of integrated information, ϕ , is an attempt to quantify a neural network’s magnitude of conscious experience. It has a long history [1, 4, 5], and at least three different measures have been called ϕ . Conceptually, the ϕ measure aims to quantify a system’s “functional irreducibility to disjoint parts”. Although innovative, the ϕ measure from [1] has some peculiarities. Using Partial Information Decomposition (PID), we derive a principled info-theoretic measure of irreducibility to disjoint parts [6]; our PID-derived measure, ψ , has numerous desirable properties over the ϕ from [1].

We aim for ψ to be a principled, well-behaved ϕ -like measure that resides purely within Shannon information theory. We compare ψ to the older ϕ measure from [1] because it is the most recent purely information-theoretic ϕ . We recognize that the most recent version of ϕ [5] knowingly and purposely sits outside standard information theory.^{1,2}

2 Preliminaries

We use the following notation throughout.

n : the number of indivisible elements in network X . $n \geq 2$.

\mathbf{P} : a partition of the n indivisible nodes clustered into m parts. Each part has at least one node and each partition has at least two parts, so $2 \leq m \leq n$.

$X_i^{\mathbf{P}}$: a random variable representing a part i at time=0. $1 \leq i \leq m$.

$Y_i^{\mathbf{P}}$: a random variable representing part i after t updates. $1 \leq i \leq m$.

X : a random variable representing the entire network at time=0. $X \equiv X_1^{\mathbf{P}} \cdots X_m^{\mathbf{P}}$.

¹The most recent version of ϕ [5] utilizes the Earth Mover’s Distance among states and thus varies with the chosen labels of the states. Although less of an issue for binary systems, a canonical property of information theories spanning from Shannon to Kolmogorov (algorithmic information theory) is invariance under relabeling of states.

²If one wished to use ψ within the larger “big phi” conceptual framework per [5] you would replace all instances of the measure “small phi” with ψ .

Y : a random variable representing the entire network after t applications of the neural network's update rule. $Y \equiv Y_1^{\mathbf{P}} \cdots Y_m^{\mathbf{P}}$.

y : a single state of the random variable Y .

\mathbf{X} : The set of the n indivisible elements at time=0.

For readers accustom to the notation in [1] the translation is: $X \equiv X_0$, $Y \equiv X_1$, $X_i^{\mathbf{P}} \equiv M_0^i$, and $Y_i^{\mathbf{P}} \equiv M_1^i$.

For pedagogical purposes we confine this paper to deterministic neural networks. Therefore all remaining entropy at time t conveys information about the past, i.e., $I(X:Y) = H(Y)$ and $I(X:Y_i^{\mathbf{P}}) = H(Y_i^{\mathbf{P}})$ where $I(\bullet:\bullet)$ is the mutual information and $H(\bullet)$ is the Shannon entropy [7]. Our model generalizes to probabilistic units with any finite number of discrete—but not continuous—states [8]. All calculations are in *bits*.

2.1 Model Assumptions

- (A) The ϕ measure is a *state-dependent* measure. Meaning that every output state $y \in Y$ has its own ϕ value. To simplify cross-system comparisons, some researchers [8] prefer to consider only the averaged ϕ , denoted $\langle \phi \rangle$. Here we adhere to the original theoretical state-dependent formulation. However, when comparing large numbers of networks we use $\langle \phi \rangle$ for convenience.
- (B) The ϕ measure aims to quantify “information intrinsic to the system”. This is often thought to be synonymous with causation, but it’s not entirely clear. But for this reason, in [1] all random variables at time=0, i.e., X and $X_1^{\mathbf{P}}, \dots, X_m^{\mathbf{P}}$ are forced to follow an *independent discrete uniform distribution*. There are actually several plausible choices for the distribution on X (see Appendix E). But for easier comparison to [1], here we also take X to be an independent discrete uniform distribution. This means that $\forall i \neq j \quad I(X_i^{\mathbf{P}}:X_j^{\mathbf{P}}) = 0$ and $H(X) = \log_2 |X|$, $H(X_i^{\mathbf{P}}) = \log_2 |X_i^{\mathbf{P}}|$ where $|\bullet|$ is the number of states in the random variable.
- (C) We set $t = 1$, meaning we compute these informational measures for a system undergoing a single update from time=0 to time=1. This has no impact on generality (see Appendix D). To analyze real biological networks one would sweep t over all reasonable timescales choosing the t that maximizes the complexity metric.

3 How ϕ Works

The ϕ measure has four steps and proceeds as follows.

1. For a given state $y \in Y$, [1] first defines the state’s *effective information* quantifying the total magnitude of information the state y conveys about X , the r.v. representing a maximally ignorant past. This turns out to be identical to [9]’s “specific-surprise”, $I(X:y)$,

$$\text{ei}(X \rightarrow y) = I(X:y) = D_{\text{KL}} \left[\Pr(X|y) \parallel \Pr(X) \right]. \quad (1)$$

Given X follows a discrete uniform distribution (assumption (B)), $\text{ei}(X \rightarrow y)$ simplifies to,

$$\begin{aligned} \text{ei}(X \rightarrow y) &= H(X) - H(X|y) \\ &= H(X) - \sum_{x \in X} \Pr(x|y) \log_2 \frac{1}{\Pr(x|y)}; \end{aligned} \quad (2)$$

in the nomenclature of [10], $\text{ei}(X \rightarrow y)$ can be understood as the “total causal power” the system exerts when transitioning into state y .

2. The second step is to quantify how much of the total causal power isn't accounted for by the disjoint parts (partition) \mathbf{P} . To do this, they define the *effective information beyond partition* \mathbf{P} ,

$$\mathbf{ei}(X \rightarrow y/\mathbf{P}) \equiv D_{\text{KL}} \left[\Pr(X|y) \left\| \prod_{i=1}^m \Pr(X_i^{\mathbf{P}}|y_i^{\mathbf{P}}) \right\| \right]. \quad (3)$$

The intuition behind $\mathbf{ei}(X \rightarrow y/\mathbf{P})$ is to quantify the amount of causal power in $\mathbf{ei}(X \rightarrow y)$ that is irreducible to the parts \mathbf{P} operating independently.³

3. After defining the causal power beyond an arbitrary partition \mathbf{P} , the third step is to find the partition that accounts for as much causal power as possible. This partition is called the *Minimum Information Partition*, or MIP. They define the MIP for a given state y as,⁴

$$\text{MIP}(y) \equiv \underset{\mathbf{P}}{\text{argmin}} \frac{\mathbf{ei}(X \rightarrow y/\mathbf{P})}{(m-1) \cdot \min_i H(X_i^{\mathbf{P}})}. \quad (4)$$

Finding the MIP of a system by brute force is incredibly computationally expensive—enumerating all partitions of n nodes scales $O(n!)$ and even for supercomputers becomes intractable for $n > 32$ nodes.

4. Fourth and finally, the system's causal irreducibility (to disjoint parts) when transitioning into state $y \in Y$, $\phi(y)$, is the effective information beyond y 's MIP,

$$\phi(y) \equiv \mathbf{ei}(X \rightarrow y/\mathbf{P} = \text{MIP}(y)). \quad (5)$$

3.1 Stateless ϕ is $\langle \phi \rangle$

Per eq. (5) ϕ is defined for every state $y \in Y$, and a single system can have wide range of ϕ -values. In [8], they found this medley of state-dependent ϕ -values unwieldy, and wanted a single number for each system. They achieved this by averaging the effective information over all states y . This results in the four corresponding stateless measures:

$$\begin{aligned} \langle \mathbf{ei}(Y) \rangle &\equiv \mathbb{E}_y \mathbf{ei}(X \rightarrow y) = I(X:Y) \\ \langle \mathbf{ei}(X \rightarrow Y/\mathbf{P}) \rangle &\equiv \mathbb{E}_y \mathbf{ei}(X \rightarrow y/\mathbf{P}) = I(X:Y) - \sum_{i=1}^m I(X_i^{\mathbf{P}}:Y_i^{\mathbf{P}}) \\ \langle \text{MIP} \rangle &\equiv \underset{\mathbf{P}}{\text{argmin}} \frac{\langle \mathbf{ei}(Y/\mathbf{P}) \rangle}{(m-1) \cdot \min_i H(X_i^{\mathbf{P}})} \\ \langle \phi \rangle &\equiv \langle \mathbf{ei}(Y/\mathbf{P} = \langle \text{MIP} \rangle) \rangle. \end{aligned} \quad (6)$$

Although the distinction has yet to affect qualitative results, researchers should note that $\langle \phi \rangle \neq \mathbb{E}_y \phi(y)$. This is because whereas each y state can have a different MIP, for $\langle \phi \rangle$ there's only one MIP for all states.

4 Three Concerns about ϕ

$\phi(y)$ **can exceed** $H(X)$. Figure 1 shows examples OR-GET and OR-XOR. On average, each looks fine—they each have $H(X) = 2$, $I(X:Y) = 1.5$, and $\langle \phi \rangle = 1.189$ bits—nothing peculiar. This changes when examining the individual states $y \in Y$.

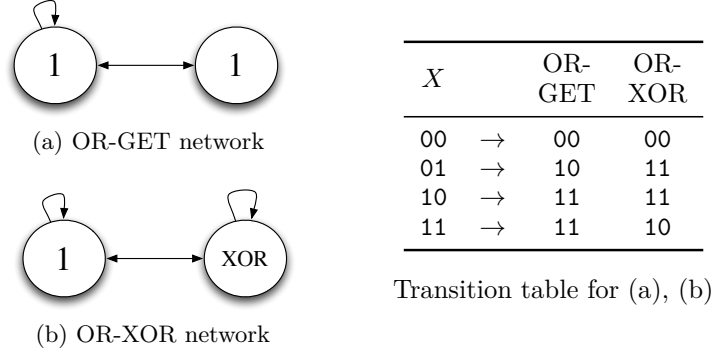
³In [1] they deviated slightly from this formulation using a process termed “perturbing the wires”. However, subsequent work [3,5] disavowed perturbing the wires and thus we don't use it here. For discussion see Appendix C.

⁴In [1] they additionally consider the *total partition* as a special case, meaning $m = 1$ and $X_1^{\mathbf{P}} = X$. However, subsequent work [3,5] disavowed the total partition and thus we don't use it here.

For OR-GET, the $\phi(y = 10) \approx 2.58$ bits. Here $\phi(y)$ *exceeds* the entropy of the entire system, $H(XY) = H(X) = 2$ bits. This means that for $y = 10$, the “irreducible causal power” exceeds not just the total causal power, $\text{ei}(X \rightarrow y)$, but ei ’s upperbound of $H(X)$! This is concerning.

For OR-XOR, $\phi(y = 11) \approx 1.08$ bits. This does not exceed $H(X)$, but it does exceed the specific surprise, $I(X : y = 11) = 1$ bit. Per eq. (6), in expectation $\langle \text{ei}(X \rightarrow Y/\mathbf{P}) \rangle \leq I(X : Y)$ for any partition \mathbf{P} . The analogous information-theoretic interpretation for a single state would be more natural if likewise $\text{ei}(X \rightarrow y/\mathbf{P}) \leq I(X : y)$ for any partition \mathbf{P} .

It’s important to note neither issue is due to normalizing in eq. (4). For OR-GET and OR-XOR there’s only one possible partition, and thus the normalization has no effect. These oddities arise from the expression for the effective information beyond a partition, eq. (3).



	OR-GET (a)				OR-XOR (b)			
	00	01	10	11	00	01	10	11
$\text{Pr}(y)$	1/4	-	1/4	1/2	1/4	-	1/4	1/2
$\text{ei}(y)$	2.00	-	2.00	1.00	2.00	-	2.00	1.00
$\phi(y)$	1.00	-	2.58	0.58	1.00	-	1.58	1.08

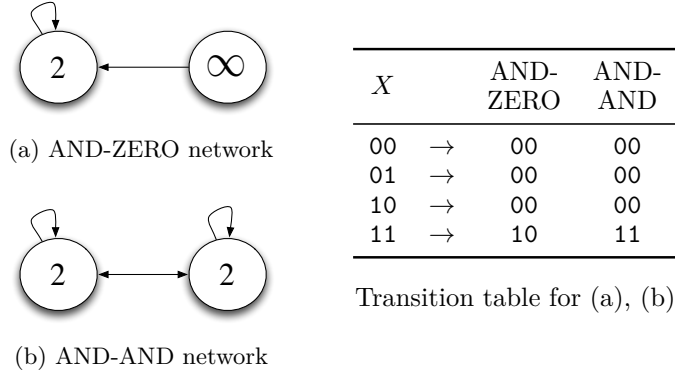
Figure 1: Example OR-GET shows that $\phi(y)$ can exceed not only $\text{ei}(X \rightarrow y)$, but $H(X)$! A dash means that particular y is unreachable for the network. The concerning ϕ values are **bolded**.

ϕ sometimes decreases with duplicate computation. In Figure 2 we take a simple system, AND-ZERO, and duplicate the AND node yielding AND-AND. We see the two systems remain exceedingly similar. Both have $H(X) = 2$ and $I(X : Y) = 0.811$ bits. Likewise, both have two Y states occurring with probability $3/4$ and $1/4$ giving $\text{ei}(X \rightarrow y)$ equal to 0.42 and 2.00 bits respectively. However, their ϕ values are quite different.

Only knowing that the ϕ ’s for AND-AND and AND-ZERO are different, we’d expect AND-AND to be higher because an AND node “does more” than a ZERO node (simply shutting off). But instead we get the opposite—AND-AND’s highest ϕ is *less* than AND-ZERO’s lowest ϕ ! The ideal measure of integrated information might be invariant or increase under duplicate computation, but it certainly wouldn’t decrease.

ϕ does not increase with cooperation among diverse parts. The ϕ measure is sometimes described as corresponding to the juxtaposition of “functional segregation” and “functional integration”. In a similar vein, ϕ is intuited as corresponding to “interdependence/cooperation among diverse parts”. Figure 3 presents four examples showing that neither intuition is well-captured by the existing ϕ measure.

In the first example, SHIFT (Figure 3a), the state of every node is shifted one-step clockwise—nothing more. The nodes are homogeneous and each node is wholly determined by its preceding node. In the three remaining networks (Figures 3b–d), every node is a function of all



	AND-ZERO (a)				AND-AND (b)			
	00	01	10	11	00	01	10	11
$\Pr(y)$	3/4	-	1/4	-	3/4	-	-	1/4
$\mathbf{ei}(y)$	0.42	-	2.00	-	0.42	-	-	2.00
$\phi(y)$	0.33	-	1.00	-	0.25	-	-	0.00

Figure 2: Examples AND-ZERO and AND-AND show that $\phi(y)$ sometimes *decreases* with duplicate computation. Here, the highest ϕ of AND-AND is *less* than the lowest ϕ of AND-ZERO. This carries into the average case with AND-ZERO’s $\langle \phi \rangle = 0.5$ and AND-AND’s $\langle \phi \rangle = 0.189$ bits. A dash means that particular y is unreachable for the network.

nodes in the network (including itself). This is to maximize the interdependence/cooperation among the nodes for high “functional integration”. Having established high cooperation, we increase the diversity or “functional segregation” from Figure 3b to 3d.

By the former intuitions, we’d expect SHIFT (Figure 3a) to have the lowest ϕ and 4321 (Figure 3d) to have the highest. But this is not the case. Instead, SHIFT, the network with the *least* cooperation (every node is a function of one other) and the *least* diverse mechanisms (all nodes have threshold 1) has a ϕ far exceeding the others—SHIFT’s lowest ϕ value at two bits dwarfs even the highest ϕ values in Figures 3b–d.

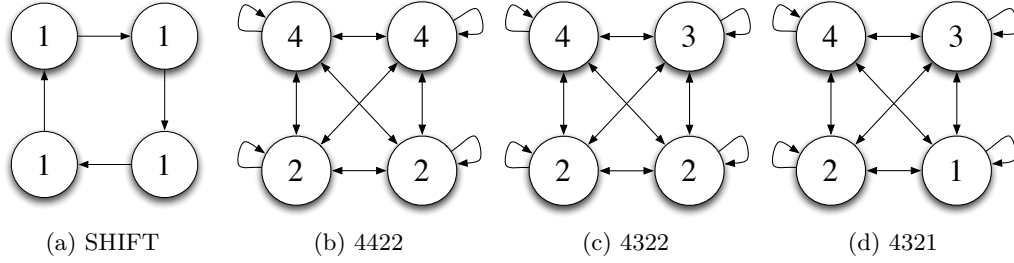
SHIFT having the highest integrated information is unexpected, but it’s not outright absurd. SHIFT does have the highest mutual information $I(X:Y)$ —so the information part is solid. Is SHIFT integrated? Well, in SHIFT each node is wholly determined by an external force (the preceding node); so SHIFT is “integrated” for a sense of the term. Whether it makes sense for SHIFT to have the highest integrated information ultimately comes down to precisely what is meant by the term “integration”. But even accepting that SHIFT is in some sense integrated, example 4321 is integrated for a palpably stronger sense of the term. Therefore, until there’s an argument that the form of integration present in SHIFT is sufficient for awareness, from a purely theoretical perspective it makes sense to prefer 4321 over SHIFT.

5 A Novel Measure of Irreducibility to a Partition

Our proposed measure ψ quantifies the magnitude of information in $I(X:y)$ (eq. (1)) that is irreducible to a partition of the system at time=0. We define our measure as,

$$\psi(\mathbf{X} : y) \equiv I(X:y) - \max_{\mathbf{P}} I_{\cup} \left(X_1^{\mathbf{P}}, \dots, X_m^{\mathbf{P}} : y \right), \quad (7)$$

where \mathbf{P} enumerates over all partitions of set \mathbf{X} , and I_{\cup} is the information about state y conveyed by the “union” across the m parts at time=0. To compute the union information I_{\cup} we use the Partial Information Decomposition (PID) framework. In PID, I_{\cup} is the



Network	$I(X:Y)$	$\min_y \phi(y)$	$\max_y \phi(y)$	$\langle \phi \rangle$
SHIFT	4.000	2.000	2.000	2.000
4422	1.198	0.000	0.673	0.424
4322	1.805	0.322	1.586	1.367
4321	2.031	0.322	1.682	1.651

Figure 3: State-dependent ϕ and $\langle \phi \rangle$ tell the same story—the ϕ value of SHIFT trounces the ϕ of the other three networks. A more intuitive complexity measure would instead increase left from to right.

inclusion–exclusion dual of I_\cap . Thus we can express I_\cup solely in terms of I_\cap by,

$$I_\cup(X_1^{\mathbf{P}}, \dots, X_m^{\mathbf{P}}:y) = \sum_{\mathbf{S} \subseteq \{X_1^{\mathbf{P}}, \dots, X_m^{\mathbf{P}}\}} (-1)^{|\mathbf{S}|+1} I_\cap(S_1, \dots, S_{|\mathbf{S}|}:y) .$$

Conceptually, the intersection information $I_\cap(S_1, \dots, S_{|\mathbf{S}|}:y)$ quantifies the magnitude of the “same information” about state y conveyed by each $S_1, \dots, S_{|\mathbf{S}|}$. Although there’s currently some debate [11,12] about what is the best I_\cap measure, there’s consensus that the intersection information n arbitrary random variables Z_1, \dots, Z_n carry about state y must satisfy the following properties:

- (**GP**) Global Positivity: $I_\cap(Z_1, \dots, Z_n:y) \geq 0$.
- (**S₀**) Weak Symmetry: $I_\cap(Z_1, \dots, Z_n:y)$ is invariant under reordering Z_1, \dots, Z_n .
- (**SR**) Self-Redundancy: $I_\cap(Z_1:y) = I(Z_1:y) = D_{\text{KL}}[\text{Pr}(Z_1|y) \parallel \text{Pr}(Z_1)]$. The intersection information a single predictor Z_1 conveys about the target state y is equal to the “specific surprise” [9].
- (**M₁**) Strong Monotonicity: $I_\cap(Z_1, \dots, Z_n, W:y) \leq I_\cap(Z_1, \dots, Z_n:y)$ with equality if there exists $Z_i \in \{Z_1, \dots, Z_n\}$ such that $I(WZ_i:y) = I(W:y)$ where WZ_i is the joint random variable (cartesian product) of W and Z_i .
- (**Eq**) Equivalence-Class Invariance: $I_\cap(Z_1, \dots, Z_n:y)$ is invariant under substituting Z_i (for any $i = 1, \dots, n$) by an informationally equivalent random variable [12].⁵ Similarly, $I_\cap(Z_1, \dots, Z_n:y)$ is invariant under substituting state y for state w if $\text{Pr}(w|y) = \text{Pr}(y|w) = 1$.

Now we take a less common course—instead of choosing a particular I_\cap that satisfies the above properties, we will simply use the properties above directly to bound the range of

⁵Meaning I_\cap is invariant under substituting Z_i with W if $H(Z_i|W) = H(W|Z_i) = 0$.

possible ψ values. Leveraging (\mathbf{M}_1) , (\mathbf{S}_0) , and (\mathbf{SR}) , eq. (7) simplifies to,⁶

$$\begin{aligned}\psi(\mathbf{X} : y) &= I(X : y) - \max_{A \subset \mathbf{X}} I_{\cup}(A, B : y) \\ &= I(X : y) - \max_{A \subset \mathbf{X}} [I(A : y) + I(B : y) - I_{\cap}(A, B : y)] ,\end{aligned}\tag{8}$$

where $A \neq \emptyset$ and $B \equiv \mathbf{X} \setminus A$.

From eq. (8), the only undefined term is $I_{\cap}(A, B : y)$. Leveraging (\mathbf{GP}) , (\mathbf{M}_1) , and (\mathbf{SR}) , we can bound it by,

$$0 \leq I_{\cap}(A, B : y) \leq \min [I(A : y), I(B : y)] .\tag{9}$$

Finally, we bound ψ by plugging in the above bounds on $I_{\cap}(A, B : y)$ into eq. (8). With some algebra and leveraging assumption (\mathbf{B}) , this yields the following bounds for ψ ,⁷

$$\begin{aligned}\psi_{\min}(\mathbf{X} : y) &= \min_{A \subset \mathbf{X}} D_{\text{KL}}[\Pr(X|y) \parallel \Pr(A|y) \Pr(B|y)] \\ \psi_{\max}(\mathbf{X} : y) &= \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X|y) \parallel \Pr(X_i) \Pr(X_{\sim i}|y)] ,\end{aligned}\tag{10}$$

where $X_{\sim i}$ is the random variable of all nodes in X excluding node i . Then, $\psi_{\min}(\mathbf{X} : y) \leq \psi(\mathbf{X} : y) \leq \psi_{\max}(\mathbf{X} : y)$.

5.1 Stateless ψ is $\langle \psi \rangle$

We define $\langle \psi \rangle$ analogous to ϕ per Section 3.1. To compute $\langle \psi \rangle$ we weaken the properties in Section 5 so that they only apply to the average case, i.e., the properties (\mathbf{GP}) , (\mathbf{M}_1) , (\mathbf{S}_0) , (\mathbf{SR}) , and (\mathbf{Eq}) don't have to apply for each $I_{\cap}(Z_1, \dots, Z_n : y)$, but merely for the average case $I_{\cap}(Z_1, \dots, Z_n : Y)$.

Via the same algebra from eq. (8), $\langle \psi \rangle$ simplifies to,

$$\begin{aligned}\langle \psi \rangle(X_1, \dots, X_n : Y) &\equiv I(X : Y) - \max_{\mathbf{P}} I_{\cup}(X_1^{\mathbf{P}}, \dots, X_n^{\mathbf{P}} : Y) \\ &= I(X : Y) - \max_{A \subset \mathbf{X}} I_{\cup}(A, B : Y) \\ &= I(X : Y) - \max_{A \subset \mathbf{X}} [I(A : Y) + I(B : Y) - I_{\cap}(A, B : Y)] ,\end{aligned}\tag{11}$$

where $A \neq \emptyset$ and $B \equiv \mathbf{X} \setminus A$. Using the weakened properties, we likewise have $0 \leq I_{\cap}(A, B : Y) \leq \min [I(A : Y), I(B : Y)]$. Plugging in these I_{\cap} bounds yields the following bounds $\langle \psi \rangle$,⁸

$$\begin{aligned}\langle \psi \rangle_{\min}(\mathbf{X} : Y) &= \min_{A \subset \mathbf{X}} I(A : B|Y) \\ \langle \psi \rangle_{\max}(\mathbf{X} : Y) &= \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X, Y) \parallel \Pr(X_{\sim i}, Y) \Pr(X_i)] ,\end{aligned}\tag{12}$$

where $X_{\sim i}$ is the random variable of all nodes in X excluding node i . Then, $\langle \psi \rangle_{\min}(\mathbf{X} : Y) \leq \langle \psi \rangle(\mathbf{X} : Y) \leq \langle \psi \rangle_{\max}(\mathbf{X} : Y)$.

6 Contrasting ψ versus ϕ

Theoretical benefits of ψ . The overarching theoretical benefit is that ψ is entrenched within the rigorous Partial Information Decomposition framework [13]. PID builds a principled irreducibility measure from a redundancy measure I_{\cap} . Here we only take the most accepted properties of I_{\cap} to bound ψ from above and below. As the complexity community converges on the additional properties I_{\cap} must satisfy [11, 12], the derived bounds on ψ will tighten.

⁶See Appendix B.1 for a proof.

⁷See Appendix B.2 for proofs.

⁸See Appendix B.3 for proofs.

There are four benefits of ψ 's principled underpinning. First, whereas $\phi(y)$ can exceed the entropy of the whole system, i.e., $\phi(y) \not\leq H(X)$, $\psi(y)$ is bounded by specific-surprise, i.e., $\psi(y) \leq I(X:y) = D_{\text{KL}}[\Pr(X|y) \parallel \Pr(X)]$. This gives ψ the natural info-theoretic interpretation for the state-dependent case which ϕ lacks. Second, PID provides justification for ψ not needing a MIP normalization and thus eliminates a longstanding ambiguity about ϕ [14]. Third, PID is a flexible framework that enables quantifying irreducibility to overlapping parts should we decide to explore it.⁹

One final perk is that ψ is already substantially faster to compute. Whereas computing ϕ scales¹⁰ $O(n!)$, computing ψ scales¹¹ $O(2^n)$ —a substantial improvement that may improve even further as the complexity community converges on additional properties of I_\cap .

Behavioral differences between ψ and ϕ . The first row in Figure 4 shows two ways a network can be irreducible to atomic elements (the nodes) yet still reducible to disjoint parts. Compare AND-ZERO (Figure 4g) to AND-ZERO+KEEP (Figure 4a). Although AND-ZERO is irreducible, AND-ZERO+KEEP reduces to the bipartition separating the AND-ZERO component and the KEEP node. This reveals how fragile measures like ψ and ϕ are—add a single disconnected node and they plummet to zero. Example 2x AND-ZERO (Figure 4b) shows that a wholly reducible network can be composed entirely of irreducible parts.

Example KEEP-KEEP (Figure 4c) highlights the only known relative drawback of ψ — ψ 's current upperbound is painfully loose.¹² The desired irreducibility for KEEP-KEEP is zero bits, and indeed, ψ_{\min} is 0 bits—but ψ_{\max} is a monstrous 1 bit! We rightly expect tighter bounds for such easy examples like KEEP-KEEP. Tighter bounds on I_\cap (and thus ψ) is an area of active research but as-is the bounds are loose.

Example GET-GET (Figure 4d) epitomizes the most striking difference between ψ and ϕ . By property (Eq), the ψ values for KEEP-KEEP and GET-GET are provably equal (making the desired ψ for GET-GET zero bits), yet their ϕ values couldn't be more different. Although the ϕ for KEEP-KEEP is zero, the ϕ for GET-GET is the maximal (!) two bits of irreducibility. Whereas ψ views GET nodes as non-integrative, ϕ views GET nodes as maximally integrative.

This begs the question—should GETs be integrative? It's sensible for GETs to be mildly integrative, but the logic of partitioning the system forces us to choose between GETs being non-integrative (akin to a KEEP) or maximally integrative. To resolve this dilemma this we return to Figure 3. The primary benefit of ψ making KEEPs and GETs equivalent is that ψ is zero for chains of GETs such as the SHIFT network (Figure 3a). This enables ψ to better match our intuition for “cooperation among diverse parts”. For example, in Figure 3 the network with the highest ϕ is the counter-intuitive SHIFT, but the network with the highest ψ is the more sensible 4321 (see table in Figure 4). With these examples in mind, we personally believe GETs being non-integrative is the better choice.

The third row in Figure 4 shows how ψ and ϕ respectively treat self-connections. In ANDtriplet (Figure 4e) and iso-ANDtriplet (Figure 4f) each node integrates information about two nodes. The only difference is that in ANDtriplet each node integrates information about two *other* nodes, while in iso-ANDtriplet each node integrates information about *itself* and one other.

Just as ψ views KEEP and GET nodes equivalently, ψ views self and cross connections equivalently. In fact, by property (Eq) the ψ values for ANDtriplet and iso-ANDtriplet are provably equal. Alternatively, ϕ considers self and cross connections differently in that ϕ can only decrease when adding a self-connection. As such, the ϕ for iso-ANDtriplet is less than ANDtriplet.

⁹Unlike disjoint parts, the maximum union information over two overlapping parts is not equal to the maximum union information over m overlapping parts. See [6] for two measures of irreducibility to overlapping parts.

¹⁰This comes from eq. (4) enumerating all partitions (Bell's number) of n elements.

¹¹This comes from eq. (8) enumerating all $2^{n-1} - 1$ bipartitions of n elements.

¹²The current upperbounds are ψ_{\max} in eq. (10) and $\langle \psi \rangle_{\max}$ in eq. (12).

The fourth row in Figure 4 shows this same self-connections business carrying over to duplicate computations. Although AND-AND (Figure 4h) and AND-ZERO (Figure 4g) perform the same computation, AND-AND has an additional self-connection that pushes AND-AND’s ϕ below that of AND-ZERO. By (**Eq**), the ψ values of AND-ZERO and AND-AND are provably equal.

7 Conclusion

Regardless of any connection to consciousness, purely as a measure of functional irreducibility we have three concerns about ϕ : (1) state-dependent ϕ can exceed the entropy of the entire system; (2) ϕ often decreases with duplicate computation; (3) ϕ doesn’t match the intuition of “cooperation among diverse parts”.

We introduced a new irreducibility measure, ψ , that solves all three concerns but otherwise stays close to the original spirit of ϕ —i.e., the quantification of a system’s irreducibility to disjoint parts. Based in Partial Information Decomposition, ψ has other desirable properties such as not needing a MIP normalization and being substantially faster to compute. We then contrasted ψ versus ϕ in binary networks.

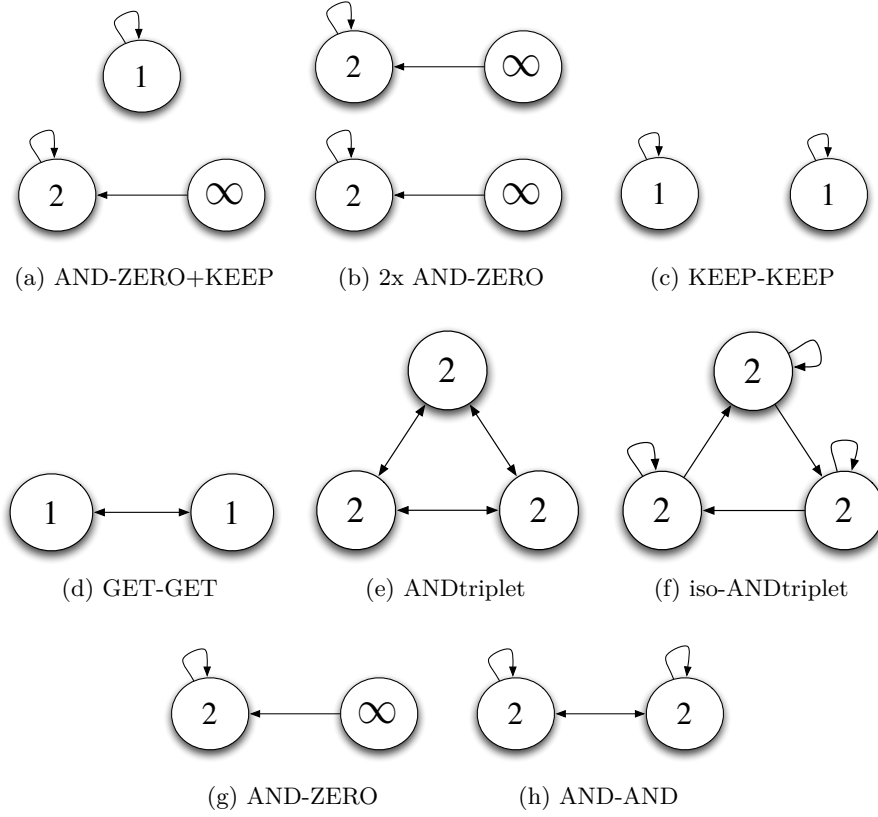
Although we endorse ψ over ϕ , the ψ measure remains imperfect. The most notable areas for improvement are:

1. The current ψ bounds are too loose. We need to tighten the I_{\cap} bounds (eq. (9)), which will tighten the derived bounds on ψ and $\langle\psi\rangle$.
2. Justify why a measure of conscious experience should prefer irreducibility to disjoint parts over irreducibility to overlapping parts.
3. Reformalize the work on qualia in [2] using ψ or comparable measure.
4. Although not specific to ψ , there needs to be a stronger justification for the chosen distribution on X (see Appendix E).

Our introduced ψ measure effortlessly generalizes to the quantum case simply by replacing all instances of Shannon mutual information in eq. (8) with von Neumann (quantum) information. This “quantum ψ ” is a quantum infotheoretic measure that remains much more faithful to its parents [1, 3] than Tegmark’s innovative perceptronium implementation [15].

References

- [1] Balduzzi D, Tononi G (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Computational Biology 4: e1000091.
- [2] Balduzzi D, Tononi G (2009) Qualia: The geometry of integrated information. PLoS Computational Biology 5.
- [3] Tononi G (2008) Consciousness as integrated information: a provisional manifesto. Biological Bulletin 215: 216–242.
- [4] Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5.
- [5] Tononi G (2012) The integrated information theory of consciousness: An updated account. Archives Italiennes de Biologie 150: 290–326.
- [6] Griffith V, Harel J (2013) Irreducibility is minimum synergy among parts. ArXiv e-prints 1311.7442.
- [7] Cover TM, Thomas JA (1991) Elements of Information Theory. New York, NY: Wiley.
- [8] Barrett AB, Seth AK (2011) Practical measures of integrated information for time-series data. PLoS Computational Biology 7.
- [9] DeWeese MR, Meister M (1999) How to measure the information gained from one symbol. Network 10: 325–340.



Network	$I(X:Y)$	$\langle\phi\rangle$	$\langle\psi\rangle_{\min}$	$\langle\psi\rangle_{\max}$
AND-ZERO+KEEP (a)	1.81	0	0	0.50
2x AND-ZERO (b)	1.62	0	0	0.50
KEEP-KEEP (c)	2.00	0	0	1.00
GET-GET (d)	2.00	2.00	0	1.00
ANDtriplet (e)	2.00	2.00	0.16	0.75
iso-ANDtriplet (f)	2.00	1.07	0.16	0.75
AND-ZERO (g)	0.81	0.50	0.19	0.50
AND-AND (h)	0.81	0.19	0.19	0.50
SHIFT (Fig. 3a)	4.00	2.00	0	1.00
4422 (Fig. 3b)	1.20	0.42	0.33	0.50
4322 (Fig. 3c)	1.81	1.37	0.68	0.88
4321 (Fig. 3d)	2.03	1.65	0.78	1.00

Figure 4: Contrasting $\langle\phi\rangle$ versus $\langle\psi\rangle$ for exemplary networks.

- [10] Korb KB, Hope LR, Nyberg EP (2009) Information-theoretic causal power. In: Information Theory and Statistical Learning, Springer. pp. 231–265.
- [11] Bertschinger N, Rauh J, Olbrich E, Jost J (2012) Shared information – new insights and problems in decomposing information in complex systems. ArXiv e-prints 1210.5902.
- [12] Griffith V, Chong EKP, James RG, Ellison CJ, Crutchfield JP (2013) Intersection information based on zero-error information and common randomness .

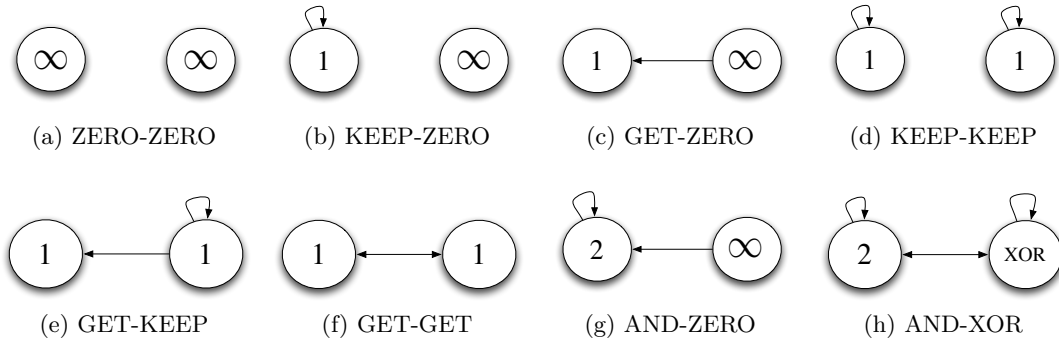
- [13] Williams PL, Beer RD (2010) Nonnegative decomposition of multivariate information. CoRR abs/1004.2515.
- [14] Balduzzi D. personal communication.
- [15] Tegmark M (2014) Consciousness as a State of Matter. ArXiv e-prints 1401.1219.
- [16] Ay N, Olbrich E, Bertschinger N, Jost J (2006) A unifying framework for complexity measures of finite systems. European Conference on Complex Systems Proceedings 2006: 202-216.
- [17] Janzing D, Balduzzi D, Grosse-Wentrup M, Schoelkopf B (2012) Quantifying causal influences. ArXiv e-prints 1203.6502.

Appendix

A Reading the Network Diagrams

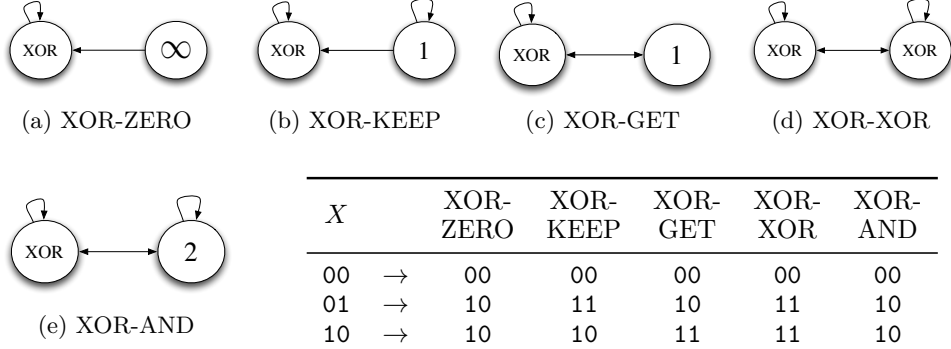
We present eight doublet networks and their transition tables so you can see how the network diagram specifies the transition table. Figure 5 shows eight network diagrams to build your intuition. The number inside each node is that node’s *activation threshold*. A node updates to 1 (conceptually an “ON”) if there at least as many of inputs ON as its activation threshold; e.g. a node with an inscribed 2 updates to a 1 if two or more incoming wires are ON. An activation threshold of ∞ means the node always updates to 0 (conceptually an “OFF”). A binary string denotes the state of the network, read left to right.

We take the AND-ZERO network (Figure 5g) as an example. Although the AND-ZERO network can never output 01 or 11 (Figure 1b), we still consider states 01, 11 as equally possible states at time=0. This is because X is uniformly distributed per assumption **(B)**.



X		ZERO-ZERO	KEEP-ZERO	GET-ZERO	KEEP-KEEP	GET-KEEP	GET-GET	AND-ZERO	AND-XOR
00	→	00	00	00	00	00	00	00	00
01	→	00	00	10	01	11	10	00	01
10	→	00	10	00	10	00	01	00	01
11	→	00	10	10	11	11	11	10	10

Figure 5: Eight doublet networks with transition tables.



X		XOR-ZERO	XOR-KEEP	XOR-GET	XOR-XOR	XOR-AND
00	→	00	00	00	00	00
01	→	10	11	10	11	10
10	→	10	10	11	11	10
11	→	00	01	01	00	01

Network	$I(X:Y)$	$\langle\phi\rangle$	$\langle\psi\rangle_{\min}$	$\langle\psi\rangle_{\max}$
ZERO-ZERO (Fig. 5a)	0	0	0	0
KEEP-ZERO (Fig. 5b)	1.0	0	0	0
KEEP-KEEP (Fig. 5d)	2.0	0	0	1.0
GET-ZERO (Fig. 5c)	1.0	1.0	0	0
GET-KEEP (Fig. 5e)	1.0	0	0	0
GET-GET (Fig. 5f)	2.0	2.0	0	1.0
AND-ZERO (Fig. 2a)	0.811	0.5	0.189	0.5
AND-KEEP	1.5	0.189	0	0.5
AND-GET	1.5	1.189	0	0.5
AND-AND (Fig. 2b)	0.811	0.189	0.189	0.5
AND-XOR (Fig. 5h)	1.5	1.189	0.5	1.0
XOR-ZERO (a)	1.0	1.0	1.0	1.0
XOR-KEEP (b)	2.0	1.0	0	1.0
XOR-GET (c)	2.0	2.0	0	1.0
XOR-AND (e)	1.5	1.189	0.5	1.0
XOR-XOR (d)	1.0	1.0	1.0	1.0

Figure 6: Networks, transition tables, and measures for the diagnostic doublets.

B Necessary Proofs

B.1 Proof that Max Union of Bipartitions Covers All Partitions

Lemma 1. *Given properties (\mathbf{S}_0) and (\mathbf{M}_1) , the maximum union information conveyed by a partition of predictors $\mathbf{X} = \{X_1, \dots, X_n\}$ about state y equals the maximum union information conveyed by a bipartition of \mathbf{X} about state y .*

Proof. We prove that the maximum information conveyed by a Partition, $\text{IcP}(\mathbf{X} : y)$, equals the maximum information conveyed by a Bipartition, $\text{IcB}(\mathbf{X} : y)$ by showing,

$$\text{IcP}(\mathbf{X} : y) \leq \text{IcB}(\mathbf{X} : y) \leq \text{IcP}(\mathbf{X} : y) . \quad (13)$$

First we show that $\text{IcB}(\mathbf{X} : y) \leq \text{IcP}(\mathbf{X} : y)$. By their definitions,

$$\begin{aligned} \text{IcP}(\mathbf{X} : y) &\equiv \max_{\mathbf{P}} \text{I}_{\cup}(\mathbf{P} : y) \\ \text{IcB}(\mathbf{X} : y) &\equiv \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \text{I}_{\cup}(\mathbf{P} : y) , \end{aligned}$$

where \mathbf{P} enumerates over all partitions of set \mathbf{X} .

By removing the restriction that $|\mathbf{P}| = 2$ from the maximization in IcB we arrive at IcP . As removing a restriction can only increase the maximum, thus $\text{IcB}(\mathbf{X} : y) \leq \text{IcP}(\mathbf{X} : y)$.

Next we show that $\text{IcP}(\mathbf{X} : y) \leq \text{IcB}(\mathbf{X} : y)$. Meaning we must show that,

$$\max_{\mathbf{P}} \text{I}_{\cup}(\mathbf{P} : y) \leq \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \text{I}_{\cup}(\mathbf{P} : y) . \quad (14)$$

Without loss of generality, we choose an arbitrary subset/part $S \subset \mathbf{X}$. This yields the bipartition of parts $\{S, \mathbf{X} \setminus S\}$. We then further partition the second part, $\mathbf{X} \setminus S$, into k (disjoint) subparts denoted T_1, \dots, T_k where $2 \leq k \leq n - |S|$ creating an arbitrary partition $\mathbf{P} = \{S, T_1, \dots, T_k\}$. We now need to show that,

$$\text{I}_{\cup}(S, T_1, \dots, T_k : y) \leq \text{I}_{\cup}(S, \mathbf{X} \setminus S : y) .$$

By (\mathbf{M}_1) equality condition, we can append each subcomponent T_1, \dots, T_k to $\{S, \mathbf{X} \setminus S\}$ without changing the union-information because for each T_i , $H(T_i | \mathbf{X} \setminus S) = 0$. Then applying (\mathbf{S}_0) we re-order the parts so that S, T_1, \dots, T_k come first. This yields,

$$\text{I}_{\cup}(S, T_1, \dots, T_k : y) \leq \text{I}_{\cup}(S, T_1, \dots, T_k, \mathbf{X} \setminus S : y) .$$

Applying (\mathbf{M}_1) inequality condition, adding the predictor $\mathbf{X} \setminus S$ can only increase the union information. Therefore we prove eq. (14), which proves eq. (13), that $\text{IcP}(\mathbf{X} : y) = \text{IcB}(\mathbf{X} : y)$. \square

B.2 Bounds on $\psi(X_1, \dots, X_n : y)$

Lemma 2. *Given (\mathbf{M}_1) , (\mathbf{SR}) and the predictors X_1, \dots, X_n are independent, i.e., $H(X) = \sum_{i=1}^n H(X_i)$, then,*

$$\psi(X_1, \dots, X_n : y) \leq \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X|y) \parallel \Pr(X_i) \Pr(X_{\sim i}|y)] .$$

Proof. Applying (\mathbf{M}_1) inequality condition, we have $I_{\cap}(A, B : y) \leq \min [I(A : y), I(B : y)]$. Via the inclusion-exclusion rule, this entails $I_{\cup}(A, B : y) \geq \max [I(A : y), I(B : y)]$, and we use this to upperbound $\psi(X_1, \dots, X_n : y)$. The random variable $A \neq \emptyset$, $B \equiv \mathbf{X} \setminus A$, and $AB \equiv X$.

$$\psi(X_1, \dots, X_n : y) = I(X : y) - \max_{A \subset \mathbf{X}} I_{\cup}(A, B : y)$$

$$\leq I(X : y) - \max_{A \subset \mathbf{X}} \max [I(A : y), I(B : y)]$$

By symmetry of complementary bipartitions, every B will be an A at some point. So we can drop the B term.

$$= I(X : y) - \max_{A \subset \mathbf{X}} I(A : y) .$$

For two parts A and A' such that $H(A|A') = 0$, $I(A : y) \leq I(A' : y)$.¹³ Therefore there will always be a maximizing subset of \mathbf{X} of size $n - 1$.

$$\begin{aligned} \psi(X_1, \dots, X_n : y) &\leq I(X : y) - \max_{\substack{A \subset \mathbf{X} \\ |A|=n-1}} I(A : y) \\ &= I(X : y) - \max_{i \in \{1, \dots, n\}} I(X_{\sim i} : y) \\ &= \min_{i \in \{1, \dots, n\}} I(X : y) - I(X_{\sim i} : y) \\ &= \min_{i \in \{1, \dots, n\}} I(X_i : y | X_{\sim i}) \\ &= \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X|y) \parallel \Pr(X_i | X_{\sim i}) \Pr(X_{\sim i} | y)] . \end{aligned}$$

Now applying that the predictors \mathbf{X} are independent, $\Pr(x_i | x_{\sim i}) = \Pr(x_i)$. This leaves,

$$\psi(X_1, \dots, X_n : y) \leq \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X|y) \parallel \Pr(X_i) \Pr(X_{\sim i} | y)] .$$

□

¹³ $I(A : y) \leq I(A' : y)$ because $I(A' : y) = I(A : y) + I(A' : y | A)$.

Lemma 3. Given **(GP)**, **(SR)** and predictors X_1, \dots, X_n are independent, i.e., $H(X) = \sum_{i=1}^n H(X_i)$, then,

$$\begin{aligned}\psi(X_1, \dots, X_n : y) &\geq \min_{A \subset \mathbf{X}} I(A : B | y) \\ &= \min_{A \subset \mathbf{X}} D_{\text{KL}} \left[\Pr(X|y) \parallel \Pr(A|y) \Pr(B|y) \right] .\end{aligned}$$

Proof. First, from the definition of I_{\cup} , $I_{\cup}(A, B : y) = I(A : y) + I(B : y) - I_{\cap}(A, B : y)$. Then applying **(GP)**, we have $I_{\cup}(A, B : y) \leq I(A : y) + I(B : y)$. We use this to lowerbound $\psi(X_1, \dots, X_n : y)$. The random variable $A \neq \emptyset$, $B \equiv \mathbf{X} \setminus A$, and $AB \equiv X$.

$$\begin{aligned}\psi(X_1, \dots, X_n : y) &= I(X : y) - \max_{A \subset \mathbf{X}} I_{\cup}(A, B : y) \\ &\geq I(X : y) - \max_{A \subset \mathbf{X}} [I(A : y) + I(B : y)] \\ &= \min_{A \subset \mathbf{X}} I(AB : y) - I(A : y) - I(B : y) \\ &= \min_{A \subset \mathbf{X}} I(A : y | B) - I(A : y) \\ &= \min_{A \subset \mathbf{X}} D_{\text{KL}} \left[\Pr(AB|y) \parallel \Pr(B|y) \Pr(A|B) \right] - D_{\text{KL}} [\Pr(A|y) \parallel \Pr(A)] \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \log \frac{\Pr(ab|y)}{\Pr(b|y) \Pr(a|b)} + \sum_a \Pr(a|y) \log \frac{\Pr(a)}{\Pr(a|y)} .\end{aligned}$$

We now add $\sum_b \Pr(b|ay)$ in front of the right-most \sum_a . We can do this because $\sum_b \Pr(b|ay) = 1.0$. Then yields,

$$\begin{aligned}\psi(X_1, \dots, X_n : y) &\geq \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \log \frac{\Pr(ab|y)}{\Pr(b|y) \Pr(a|b)} + \Pr(b|ay) \Pr(a|y) \log \frac{\Pr(a)}{\Pr(a|y)} \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \left[\log \frac{\Pr(ab|y)}{\Pr(b|y) \Pr(a|b)} + \log \frac{\Pr(a)}{\Pr(a|y)} \right] \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \log \frac{\Pr(ab|y) \Pr(a)}{\Pr(a|y) \Pr(b|y) \Pr(a|b)} .\end{aligned}$$

Now applying that the predictors \mathbf{X} are independent, $\Pr(a|b) = \Pr(a)$; thus we can cancel $\Pr(a)$ for $\Pr(a|b)$. This yields,

$$\begin{aligned}\psi(X_1, \dots, X_n : y) &\geq \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \log \frac{\Pr(ab|y)}{\Pr(a|y) \Pr(b|y)} \\ &= \min_{A \subset \mathbf{X}} D_{\text{KL}} \left[\Pr(X|y) \parallel \Pr(A|y) \Pr(B|y) \right] .\end{aligned}$$

□

B.3 Bounds on $\langle \psi \rangle(X_1, \dots, X_n : Y)$

Lemma 4. *Given (\mathbf{M}_1) , (\mathbf{SR}) and the predictors X_1, \dots, X_n are independent, i.e., $H(X) = \sum_{i=1}^n H(X_i)$, then,*

$$\langle \psi \rangle(X_1, \dots, X_n : Y) \leq \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X, Y) \parallel \Pr(X_{\sim i}, Y) \Pr(X_i)] .$$

Proof. First, using the same reasoning in Lemma 2, we have,

$$\begin{aligned} \langle \psi \rangle(\mathbf{X} : Y) &\leq I(X : Y) - \max_{i \in \{1, \dots, n\}} I(X_{\sim i} : Y) \\ &= \min_{i \in \{1, \dots, n\}} I(X : Y) - I(X_{\sim i} : Y) \\ &= \min_{i \in \{1, \dots, n\}} I(X_i : Y | X_{\sim i}) \\ &= \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X, Y) \parallel \Pr(X_i | X_{\sim i}) \Pr(X_{\sim i}, Y)] . \end{aligned}$$

Now applying that the predictors \mathbf{X} are independent, $\Pr(X_i | X_{\sim i}) = \Pr(X_i)$. This yields,

$$\langle \psi \rangle(\mathbf{X} : Y) \leq \min_{i \in \{1, \dots, n\}} D_{\text{KL}}[\Pr(X, Y) \parallel \Pr(X_{\sim i}, Y) \Pr(X_i)] .$$

□

Lemma 5. *Given (\mathbf{GP}) , (\mathbf{SR}) and predictors X_1, \dots, X_n are independent, i.e., $H(X) = \sum_{i=1}^n H(X_i)$, then,*

$$\langle \psi \rangle(X_1, \dots, X_n : Y) \geq \min_{A \subset \mathbf{X}} I(A : B | Y) .$$

Proof. First, using the same reasoning in Lemma 3, we have,

$$\begin{aligned} \langle \psi \rangle(X_1, \dots, X_n : Y) &\geq I(X : Y) - \max_{A \subset \mathbf{X}} [I(A : Y) + I(B : Y)] \\ &= \min_{A \subset \mathbf{X}} I(AB : Y) - I(A : Y) - I(B : Y) \\ &= \min_{A \subset \mathbf{X}} I(A : B | Y) - I(A : B) . \end{aligned}$$

Now applying that the predictors \mathbf{X} are independent, $I(A : B) = 0$. This yields,

$$\langle \psi \rangle(X_1, \dots, X_n : Y) \geq \min_{A \subset \mathbf{X}} I(A : B | Y) .$$

□

C Definition of Intrinsic ei a.k.a. “Perturbing the Wires”

State-dependent ei across a partition, $\text{ei}(X \rightarrow y/\mathbf{P})$, is defined by eq. (15).

$$\begin{aligned} \text{ei}(X \rightarrow y/\mathbf{P}) &\equiv D_{\text{KL}} \left[\Pr(X \rightarrow y) \left\| \prod_{i=1}^m \Pr(X_i^{\mathbf{P}} \rightarrow y_i^{\mathbf{P}}) \right\| \right] \\ &= D_{\text{KL}} \left[\Pr(X|y) \left\| \prod_{i=1}^m \Pr^*(X_i^{\mathbf{P}}|y_i^{\mathbf{P}}) \right\| \right]. \end{aligned} \quad (15)$$

Balduzzi/Tononi [1] define the probability distribution describing the intrinsic information from the whole system X to state y as,

$$\Pr(X \rightarrow y) = \Pr(X|y) = \left\{ \Pr(x|y) : \forall x \in X \right\}.$$

They then define probability distribution describing the intrinsic information from a part $X_i^{\mathbf{P}}$ to a state $y_i^{\mathbf{P}}$ as,

$$\Pr^*(X_i^{\mathbf{P}} \rightarrow y_i^{\mathbf{P}}) \equiv \Pr^*(X_i^{\mathbf{P}}|y_i^{\mathbf{P}}) = \left\{ \Pr^*(x_i^{\mathbf{P}}|y_i^{\mathbf{P}}) : \forall x_i^{\mathbf{P}} \in X_i^{\mathbf{P}} \right\}.$$

First we define the fundamental property of the \Pr^* distribution.¹⁴ Given a state $x_i^{\mathbf{P}}$, the probability of a state $y_i^{\mathbf{P}}$ is computed by probability each node in the state $y_i^{\mathbf{P}}$ independently reaches the state specified by $y_i^{\mathbf{P}}$,

$$\Pr^*(y_i^{\mathbf{P}}|x_i^{\mathbf{P}}) \equiv \prod_{j=1}^{|\mathbf{P}_i|} \Pr(y_{i,j}^{\mathbf{P}}|x_i^{\mathbf{P}}). \quad (16)$$

Then we define the join distribution relative to eq. (16):

$$\Pr^*(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}) = \Pr^*(x_i^{\mathbf{P}}) \Pr^*(y_i^{\mathbf{P}}|x_i^{\mathbf{P}}) = \Pr^*(x_i^{\mathbf{P}}) \prod_{j=1}^{|\mathbf{P}_i|} \Pr(y_{i,j}^{\mathbf{P}}|x_i^{\mathbf{P}}).$$

Then applying assumption **(B)**, X follows a discrete uniform distribution, so $\Pr^*(x_i^{\mathbf{P}}) \equiv \Pr(x_i^{\mathbf{P}}) = 1/|X_i^{\mathbf{P}}|$. This gives us the complete definition of $\Pr^*(x_i^{\mathbf{P}}, y_i^{\mathbf{P}})$,

$$\Pr^*(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}) = \Pr(x_i^{\mathbf{P}}) \prod_{j=1}^{|\mathbf{P}_i|} \Pr(y_{i,j}^{\mathbf{P}}|x_i^{\mathbf{P}}). \quad (17)$$

With the joint \Pr^* distribution defined, we can compute anything we want—such as the expressions for $\Pr^*(y_i^{\mathbf{P}})$ and $\Pr^*(x_i^{\mathbf{P}}|y_i^{\mathbf{P}})$ —by summing over eq. (17),

$$\Pr^*(y_i^{\mathbf{P}}) = \sum_{x_i^{\mathbf{P}} \in X_i^{\mathbf{P}}} \Pr^*(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}) \quad (18)$$

$$\Pr^*(x_i^{\mathbf{P}}|y_i^{\mathbf{P}}) = \frac{\Pr^*(x_i^{\mathbf{P}}, y_i^{\mathbf{P}})}{\Pr^*(y_i^{\mathbf{P}})}. \quad (19)$$

¹⁴It's worth noting that $\Pr^*(X|y) \neq \Pr(X|y)$.

D Setting $t = 1$ Without Loss of Generality

Given t stationary surjective functions that may be different or the same, denoted $f_1 \cdots f_t$, we define the state of system at time t , denoted X_t , as the application of the t functions to the state of the system at time 0, denoted X ,

$$X_t = f_t \left(f_{t-1} \left(\cdots f_2 (f_1 (X)) \cdots \right) \right) .$$

We instantiate an empty “dictionary function” $g(\bullet)$. Then for every $x_0 \in X_0$ we assign,

$$g(x) \equiv f_t \left(f_{t-1} \left(\cdots f_2 (f_1 (x)) \cdots \right) \right) .$$

At the end of this process we have a function g that accomplishes any chain of stationary functions $f_1 \cdots f_t$ in a single step for the entire domain X . So instead of studying the transformation,

$$X \xrightarrow{f_1 \cdots f_t} X_t ,$$

we can equivalently study the transformation,

$$X \xrightarrow{g} Y .$$

Here’s an example using mechanism $f_1 = f_2 = f_3 = f_4 = \text{AND-GET}$.

time=0		$t = 1$		$t = 2$		$t = 3$		$t = 4$
00	→	00	→	00	→	00	→	00
01	→	00	→	00	→	00	→	00
10	→	01	→	00	→	00	→	00
11	→	11	→	11	→	10	→	00
$g(\bullet)$		AND-GET		AND-AND		AND-ZERO		ZERO-ZERO

Table 1: Applying the update rule “AND-GET”, over four timesteps.

E The Appropriate Distribution on X is Ambiguous

A system’s “mechanism” is defined by the probability distribution $\Pr(Y|X)$. And we are asking that given a state $Y = y$, how clearly are the possible states of X specified—i.e., Given the mechanism $\Pr(Y|X)$, how different are the distributions $\Pr(X)$ and $\Pr(X|y)$? To compute $\Pr(X|y)$ from $\Pr(Y|X)$, we must define a distribution $\Pr(X)$. There are several choices for $\Pr(X)$. These are some of the prominent ones:

Empirical: Make X follow the distribution actually recorded from the system.

Discrete uniform: Every state $x \in X$ has $\Pr(x) = \frac{1}{|X|}$ where $|X|$ is the number of distinct states of r.v. X .

Capacity: Regardless of state $y \in Y$, the X distribution is,

$$X \sim \underset{\Pr(X')}{\operatorname{argmax}} \operatorname{I}(X':Y)$$

Each of these distributions have been used for causal measures [10, 16, 17]. And for each of these candidate distributions on X , there exist (causal) questions for which it is the best/most appropriate choice. Therefore, merely saying we want a “causal measure” for conscious experience does not rule any of them out. Conceptually, it makes sense to preclude the empirical distribution as it does not take into account counterfactuals. But what about the discrete-uniform versus the capacity distribution? What reason is there to prefer one over the other? Ideally this would be answered by returning to the original thought experiments for consciousness.